

DPaper：一种面向语义出版的结构化论文写作工具设计与实现

乐小虬[1] 王子璇[1,2] 张晓林[1] 何远标[1] 付常雷[1] 许丽媛[1]
[1](中国科学院文献情报中心 北京 100190)
[2](中国科学院大学 北京 100049)

摘要:【目的】面向语义出版构建论文写作工具，在论文写作阶段实现内容结构化、对象化，使得一篇论文即是一个系统，论文可运行、可交互、可体验。【方法】采用数字对象和数字模板技术将论文内容(元数据、章节、数据、富媒体等)分解成不同类型数字对象，数字对象间采用模板进行组织，通过事件触发机制实现交互，采用HTML5网页形式进行编辑和呈现并存储为XML结构化文档包。【结果】DPaper结构化论文写作工具已上线，提供从素材收集(云笔记)、数字对象制作、自动标引参考文献、按期刊版式呈现到Word文档格式转换等一系列功能，论文内容实现对象化和部分语义化。【局限】与常规论文编辑器相比，数字对象编辑器功能还不完善，还不能创建公式、图形等对象，排版的灵活性不足。【结论】利用DPaper写作工具可以在写作阶段由作者构建出满足语义出版应用需求的结构化论文。

关键词: DPaper 语义出版 结构化论文 数字对象 写作工具
分类号: TP393

1 引 言

数字环境下学术论文的形态及利用方式出现了许多新的趋势，内容结构化、对象化、语义化等数字化表征使得面向细粒度内容的应用不断推陈出新。语义出版作为一种新的应用形式近年来进入快速发展时期，它能使论文间的数据整合变得更加容易[1]，知识对象与知识关系可进行鉴别和标引，并把解析逻辑与结果作为内容出版的有机组成部分[2]。STM 2015年技术趋势报告认为[3]：期刊论文正处于“Hub and Spoke”出版模型的中心位置，连接视频、图形、表格，以及多种不同的数字对象(Artifacts)，其中数据正上升为首要(First-Class)研究对象。在其2014年的趋势图中描绘了新式论文的场景[4]：可计算、富集化(含可交互的数据查看器、图形/图像、图表、可视化、活的方程等)，可使出版变成软件模型，作者和研究者均具有更好的体验。

然而，目前还没有一种真正可用的论文写作工具，辅助作者在创作阶段生成满足语义出版要求的结构化论文。为此，笔者开发了一种面向语义出版的结构化论文写作工具——DPaper (http://idpaper.las.ac.cn/)，旨在从根本上改变论文的利用模式，在写作阶段实现内容结构化、对象化，论文变成软件模型，一篇论文即是一个系统，论文可运行、可交互、可体验，作者的研究数据、研究过程及研究结果可为读者操作和复用。本文将阐述该工具主要研究思路和系统设计实现方法。

通讯作者：乐小虬，ORCID：0000-0002-7114-5544，E-mail：lexq@mail.las.ac.cn。
*本文部分内容已发表于《Journal of Data and Information Science》。

2　相关研究

近年来，在面向语义出版的结构化论文研究方面已有一些探索，基本思路是使论文内容对象化和语义化。主要有三种做法：将论文内容模块化；数字对象封装并进行语义描述；语义标注。具有代表性的是模块化论文模型和语义出版模型。语义标引是目前文献内容语义化的主要手段，有大量研究与试验[2]，本文不作赘述。

模块化论文模型由 Kircz 提出[5-6]，论文由模块组成，模块被定义为具有独一无二特性、自含概念表示的信息单元，数据集、图像、音频、视频等被看成是独立但可交互的对象或模块聚合到论文中，为便于交流，模块被连接成固定单元。模块化结构能给阅读和出版带来更高效率，*Cell* 中的论文使用了这种模块结构。利用数字对象组织学位论文也是模块化思想的典型应用，做法是将数字对象应用融入现有的电子化学位论文系统中，提供 METS/XML 转换、导入导出功能，典型工具为 OpenETD，它既是一个独立的学位论文提交系统，也是一个利用 METS/XML 导出功能实现机构仓储的组件[7]。在 ProQuest/UMI 系统中，音频、视频、数据集(SpreadSheets)等富媒体均以学位论文补充文件的形式在线提交到系统中，提交时需填写相应的描述信息[8]。

在语义出版模型中，Hunter 提出科学出版包(Scientific Publication Package, SPP)新信息格式[9]，用于封装原始数据、来源产品、算法、软件、文本、相关上下文环境以及原数据，使科学家能够获取、索引、存储、共享、交换、重用、比较和集成科学结果。SPP 基于许多科学概念模型，是一种用 RDF 包表示的复合数字对象，复合对象内部原子对象间的关系要么在元数据获取时从本体规则推理中明确定义，要么由科学家在 SPP 描述时定义。强调工作流技术作为科学过程的组成部分，用于获取产生科学数据和来源产品的处理步骤链，可使科学家们以一种可重复、可证实、分布的方式描述和执行他们的实验过程，追踪错误来源、处理缺陷[10]。

在面向语义出版的结构化论文编辑工具方面，目前还没有看到通用性工具，但 BioLit 项目和 SCOPE (Scientific Compound Object Publishing and Editing System)项目分别从语义标记和复合数字对象的角度开展了有益探索。Fink 等在 BioLit 项目中开发了基于 XML 的写作工具，利用美国医学图书馆的文档类型定义(NLM DTD)存储标准化且机器可读的出版物[11]。这个 DTD 也包括一些对文章本身和对象内容(如图、表)的语义标记、唯一标识符。该工具将为开放文献和生物学数据的集成提供方便，使用 PLoS 和 Protein Data Bank (PDB)的全部语料做测试[10]。

SCOPE 工具开展了让研究者自己构建数字对象的尝试，这是面向语义出版结构化论文的最终出路，因为只有作者最清楚具体的研究过程、计算方法、实验材料、实验数据及结果[12]。SCOPE 是一个利用 OAI-ORE 规范的数字内容串联工具，它是一种科学复合对象出版和编辑系统，设计用于使科学家易于创作、出版和编辑科学复合对象，使科学家封装科学实验或发现过程中不同的数据集和资源，单个复合对象可出版和交换[12]。但 SCOPE 构造示例显示复合对象的构造过程需借助语义网关系，即使是 ICT 专家也很难完成构建，无论是可用性还是实用性均很难满足现实需求。

3　系统设计

3.1　面向语义出版的论文写作工具概念模型

目前论文写作大多采用文档编辑器(如 Word 等)，论文内容以静态复合文档(如 Doc/PDF 格式)的形式存在，内容非结构化、非语义化、且交互能力弱，使得作者的研究数据和过程较难完整、有效地展示给读者，研究结果不易被同行有效利用、理解、观察和验证。

面向语义出版的论文写作工具是一种可计算、可复用/验证、可交互的论文系统，其内容可操作、可组配、可发布，具备多种呈现方式。利用这种工具创作的论文应具备以下能力:

(1) 论文可运行，一篇论文可发布成一个应用系统;

(2) 论文用数字对象表示，内容结构化、语义化;

(3) 具有丰富的富媒体对象，能充分展示科学研究过程及成果;

(4) 数字对象可独立运行，可组合、定制，满足语义出版的需求。

DPaper 在设计中引入模块化论文模型的思想[5]，

通过构建规范化数字论文模板，将学位论文的内容组织与表现分离，学术论文不再是一个静态复合文档，而是一种可配置、可操作、可传递、可交换、可保存的数字对象集。数字对象间彼此关联，内容的组织采用开放的元数据标准(如 METS、Dublin Core 等)进行描述，内容的呈现采用 Web 形式进行展示。数字对象的操作上，制定相应处理规范及接口标准，将一些复合数字对象变成可集成的微服务。

3.2 系统处理框架

DPaper 系统框架如图 1 所示:

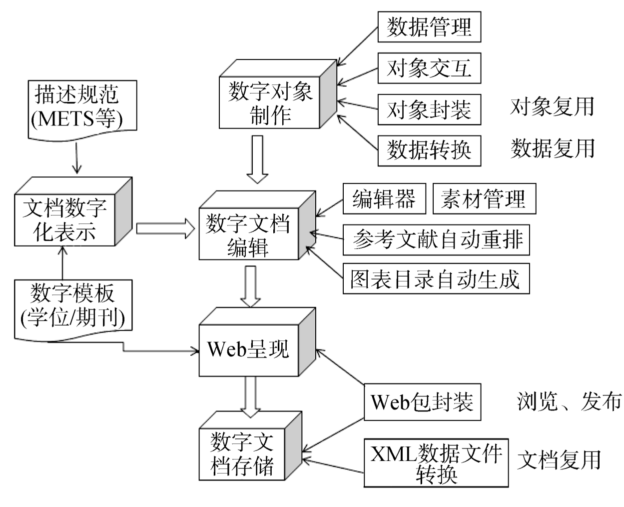


图 1 DPaper 系统处理框架

DPaper 系统由文档数字化表示、数字对象制作、数字文档编辑、文档 Web 呈现以及数字文档存储 5 部分构成:

(1) 文档数字化表示主要是利用描述规范制定相应的数字化论文模板;

(2) 数字对象制作负责对象数据的管理、对象交互、封装以及数据的转换等处理;

(3) 数字文档编辑是数字论文创作、编辑、修改的场所，以论文组织结构单元为基础组织数字对象，在编辑过程中数字对象被赋予语义标签，从而实现对象的结构化和语义化;

(4) 编辑器中的数字对象按照模板进行个性化组配，彼此间数据关联，以网页的形式呈现，提供浏览和发布功能;

(5) 论文中的数字对象及其数据以 Web 包的形式进行存储，同时将结构描述信息存储于 XML 数据文件中，用于数字文档的交换和第三方软件复用。

4 关键技术方法

4.1 论文数字对象组织

对论文内容按粒度进行分解和描述，参照 METS、Dublin Core 以及 NCBI 和 NLM 制定的图书与收藏标签库(Book and Collection Tag Library version 3.0)[13]对学位论文内容进行规范化描述，如图 2 所示:

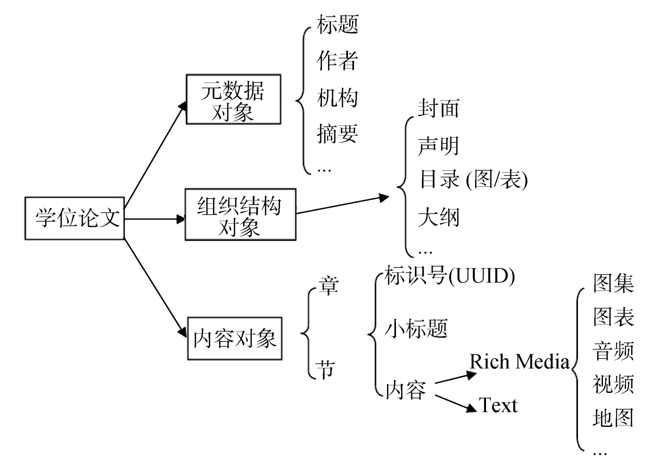


图 2 结构化论文数字对象组织框架

描述框架由论文元数据对象、论文组织结构对象和内容对象三部分构成:

(1) 论文元数据采用 Dublin Core 中的元素集描述，包括论文标题、作者、机构、中文摘要、英文摘要、关键词等;

(2) 论文组织结构对象涵盖封面、声明、目录、大纲、参考文献、致谢、附录等类别;

(3) 内容对象包括章、小节；富媒体对象涵盖图表、图片/图集、音频、视频、动画、地图、数据、网络图、算法、软件等。每个数字对象用 UUID 进行统一标识。例如，章节对象(Section)记录每章的结构和对应内容，使用 Sec-id(章节编号), Sec-title(章节标题), Sec-content(章节内容)三个元素进行描述，图片/图集、图表、音频、视频等数据均以对象的形式进行描述并嵌入相应的章节中。

4.2 数字对象间通信与交互

Dpaper 系统中的数字对象集数据载入、处理、编辑、呈现、存储等操作为一体，数字对象之间数据可以相互调用，一个对象的数据操作可即时触发另一个对象在编辑、运行中的状态和结果。这种交互过程主要通过事件触发机制完成，当用户执行某些操作时，

系统内部会通知相应数字对象做出更新响应。

在 DPaper 中，数字对象定义了多种响应事件，主要包括：新增、开始修改、结束修改、删除、复制、擦除等。事件处理过程：文档操作(按钮/快捷键)→环境处理→触发 Before 事件，做权限检查等相关操作→触发 Manager 响应事件→触发 After 事件→完成事件。数据对象基本事件(新增、修改、删除)响应流程如图 3 所示：

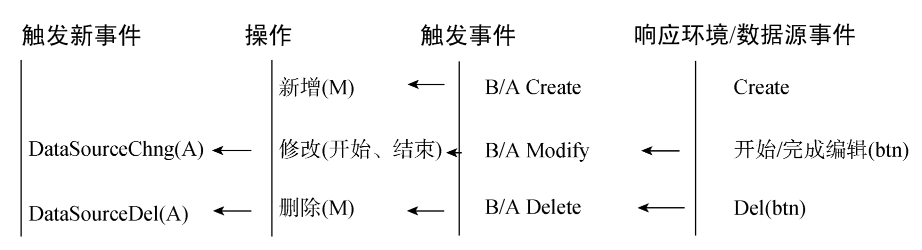


图 3　数字对象事件响应流程

4.3　数字对象语义描述

采用数字模板技术对数字论文中定义的不同粒度的数字对象标记相应语义标签，由系统自动生成。为了减少编辑操纵的复杂度，系统暂未对数字对象中的内容进行更为详细的语义化处理。在 DPaper 中，选用“中国科学院硕士学位论文”模板以及《现代图书情报技术》期刊论文模板构建了学位论文和期刊论文两类数字化模板。模板采用 4.1 节中定义的元素进行标记，论文中不同粒度的数字对象(如封面、声明、标题、作者、机构、章节、图表、参考文献等)的字体、字号、样式、位置、构成、格式在模板做了具体定义，以 XML 文件的形式记录在系统中，供论文对象编辑时调用。图 4、图 5 分别是学位论文封面对象和参考文献对象的语义描述实例。


```
<?xml version="1.0" ?>
- <iDisseration>
 - <cover>
    <Classification>G250</Classification>
    <Secret>XXX</Secret>
    <Colleage>中国科学院</Colleage>
    <ArticleType>硕士学位论文</ArticleType>
    <ArticleTitle>学术型网络数据库功能发展趋势与评价体系研究</ArticleTitle>
    <ArticleAuthor>丁艳君</ArticleAuthor>
    <GraduateTeacher>张硕</GraduateTeacher>
    <GraduateTeacherCompany>中国科学院文献情报中心</GraduateTeacherCompany>
    <DegreeType>硕士学位</DegreeType>
    <Major>图书馆学</Major>
    <ArticleTime>2009年12月</ArticleTime>
    <ArticleTimeDefence>2010年1月</ArticleTimeDefence>
    <CultivateCollege>中国科学院文献情报中心</CultivateCollege>
   </cover>
 - <coverE>
    <ArticleTitleE>The Study on Trend and Evaluation System of Resources Database Functions</ArticleTitleE>
    <ArticleTypeE>Doctoral/Master</ArticleTypeE>
    <ArticleType2E>Dissertation/Thesis</ArticleType2E>
    <MajorE>Management in the Library</MajorE>
    <ArticleAuthorE>Ding YanJun</ArticleAuthorE>
    <GraduateTeacherE>Zheng JianCheng</GraduateTeacherE>
    <ArticleTimeE>January,2010</ArticleTimeE>
   </coverE>
 - <Statement>
    <CollegeS>中国科学院文献情报中心</CollegeS>
    <AuthorS>丁艳君</AuthorS>
    <TeacherS>王企鹅额</TeacherS>
    <TimeS>2010.1</TimeS>
   </Statement>
 - <Abstract>
```


图 4　学位论文封面对象语义描述

4.4　学位论文数字对象重用机制

DPaper 目前提供数据重用、数字对象重用以及整篇论文重用三种模式。



图 5　学位论文参考文献对象语义描述

(1) 数据重用

数字对象中的数据通过格式转换达到数据复用的目的。Dtable、Dchart、关系图等数字对象中的数据可转换成 JSON、CSV、RDF/XML 等格式数据文件。

(2) 数字对象的重用

数字对象中的数据、程序及库文件等封装成独立的 Web 包，具有访问入口和对象描述元数据，离开 DPaper 环境能独立在浏览器中运行。对象复制后，能嵌入其他数字论文中复用。

(3) 整篇论文重用

在系统内部，DPaper 使用论文结构、论文数据、显示格式三个 XML 文件：论文结构文件用于记录论文对象间的层级关系；数据文件记录论文的元数据以及数字对象的内容、路径、位置等信息；显示格式文件记录论文中各类数字对象在规范显示模板中的显示信息，如字体、字号、位置、颜色、样式等。这些内部结构文件在用于数字保存或文档交换时，采用分离式 METS 电子文档的形式进行转换，如图 6 所示。

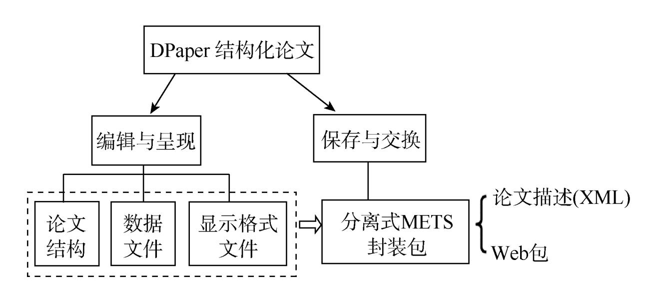


图 6 DPaper 结构化论文的存储与转换

5 工具应用

5.1 DPaper 软件构成及主要功能

DPaper 由论文编辑器、云笔记和 Word 插件三部分构成，如图 7 所示。

各部分主要功能如下:

(1) 论文编辑器是 PC 端桌面软件，为结构化论文构建的主平台，负责论文的创建、数字对象的创建、制作，对象内容的编辑、数据管理、对象管理、Web 预览、文档转换等功能;

(2) 云笔记用于收集论文研究素材及团队协作写作，可在 PC 端、手机端、网页端摘录笔记并实现数据同步;

(3) Word 插件用于在 Word 环境下构建 DPaper 结构化文档，使得两种文档间实现转换。

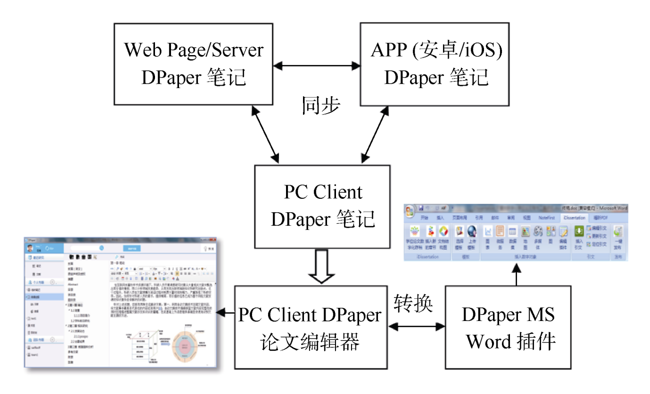
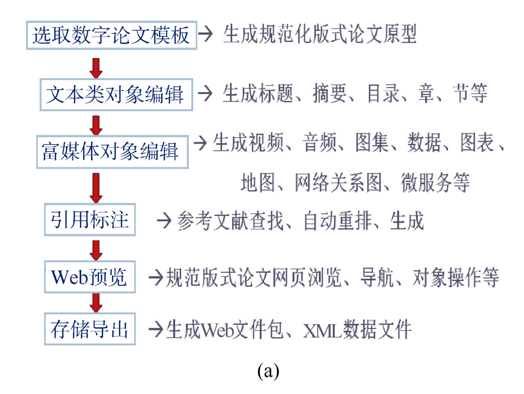


图 7 DPaper 软件构成

5.2 DPaper 论文构建过程

DPaper 结构化论文构建流程如图 8(a)所示。图 8(b)是利用 DPaper 生成的一篇中国科学院硕士学位论文结果界面截图。



(a)



(b)

图 8 DPaper 结构化硕士学位论文构建实例

6 结 语

DPaper 探索了面向语义出版的结构化论文写作工具的构建方法，面向论文写作生命周期设计实现了一套实用软件系统。采用数字模板机制将论文表示为数据、结构、呈现样式三种相互分离又彼此关联的结构化文件，在写作阶段实现论文内容的结构化和部分语义化。富媒体对象的引入使得论文具备较高的可操作性和可复用性，论文的呈现可根据具体应用分别产生 Word 文档、可为机器阅读的 XML 文档、可运行的 Web 文档系统。从而在创作的源头上实现论文的结构化和部分语义化，对于推动语义出版具有重要意义。

DPaper 还有一些不足：主平台常规的编辑功能与 Word 相比还有较大差距，还不能创建公式、图形等对象，对于长期习惯于用 Word、WPS 撰写论文的作者来说，用户能否习惯这种新的编辑模式还有待实践验证。另外数字对象在主平台和 Word 插件间还不能平滑调用，稳定性需进一步提高。后期将针对这些问题对系统做相应改进。

参考文献:

[1] Shotton D. Semantic Publishing: The Coming Revolution in

Scientific Journal Publishing [J]. Learned Publishing, 2009, 22(2): 85-94.

[2] 张晓林. 颠覆数字图书馆的大趋势[J]. 中国图书馆学报, 2011, 37(5): 4-12. (Zhang Xiaolin. The Forces Disrupting Digital Library [J]. Journal of the Library Science in China, 2011, 37(5): 4-12.)

[3] From STM, Tech Trends for 2015 [EB/OL]. [2016-10-11]. http://beyondthebookcast.com/from-stm-tech-trends-for-2015/.

[4] STM Tech Trends 2014 [EB/OL]. [2016-10-11]. http://www.stm-assoc.org/2014_04_29_Innovations_USA_STM_Tech_Trends_2014.pdf.

[5] Kircz J G. Modularity: The Next Form of Scientific Information Presentation? [J]. Journal of Documentation, 1998, 54(2): 210-235.

[6] Kircz J G. New Practices for Electronic Publishing 2: New Forms of the Scientific Paper [J]. Learned Publishing, 2002, 15(1): 27-32.

[7] OpenETD: Open Source Electronic Theses and Dissertations Managment Software [DB/OL]. [2015-08-06]. https://rucore.libraries.rutgers.edu/open/projects/ openetd/ index.php.

[8] ProQuest Dissertation Publishing [DB/OL]. [2015-08-06]. http://www.etdadmin.com/UMI_PreparingYourManuscriptGuide.pdf.

[9] Hunter J. Scientific Publication Packages-A Selective Approach to the Communication and Archival of Scientific Output [J]. Journal of Digital Curation, 2006, 1(1): 3-16.

[10] Enhanced Publications [EB/OL]. [2016-10-11]. http://www.doc88.com/p-873117284280.html.

[11] Fink J L, Bourne P E. Reinventing Scholarly Communication for the Electronic Age [J]. CTWatch Quarterly, 2007, 3(3): 26-31.

[12] Cheung K, Hunter J, Lashtabeg A, et al. SCOPE: A Scientific Compound Object Publishing and Editing System [J]. International Journal of Digital Curation, 2008, 3(2): 4-18.

[13] Book and Collection Tag Library Version 3.0 [EB/OL]. [2016-10-11]. http://dtd.nlm.nih.gov/book/tag-library/3.0/index.html.

作者贡献声明:

乐小虬: 提出面向语义出版结构化论文研究思路, 设计研究方案, 负责系统实施, 论文撰写与修改;
王子璇: 文献调研, 软件开发、测试, 论文修改;
张晓林: 提出 Smart Dissertation 研究方向和目标, 论文修改;
何远标: 负责系统开发, 解决关键问题;
付常雷, 许丽媛: 模块开发、测试。

利益冲突声明:

所有作者声明不存在利益冲突关系。

收稿日期: 2016-09-13
收修改稿日期: 2016-10-19

DPaper: A Structured Paper Authoring Tool for Semantic Publishing

Le Xiaoqiu[1] Wang Zixuan[1,2] Zhang Xiaolin[1] He Yuanbiao[1] Fu Changlei[1] Xu Liyuan[1]
[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)
[2] (University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: [**Objective**] We developed a paper authoring tool for semantic publishing, which makes the article's content structured and object-oriented. Each paper is a system with executable, interactive and experiential features. [**Methods**] First, we divided the content of each paper (metadata, chapters, data, media etc.) into objects organized by digital template. Second, these elements interacted with each other through the event trigger mechanism. Finally, the paper was modified and presented with HTML5 pages, and then, saved as XML documents. [**Results**] DPaper is available at iDPaper.las.ac.cn, which provides a series of functions such as material collection (cloud notes), digital object creation, automatic reference indexing, Word document format conversion in accordance with periodical layouts etc. The paper's content is object oriented and partial semantization. [**Limitations**] Compared to conventional paper editors, the DPaper's digital object editor could not create formulas or graphics, and is not flexible to change layouts. [**Conclusions**] DPaper could help us compose a structured paper that meets the requirements of semantic publishing.

Keywords: DPaper Semantic publishing Structured paper Digital object Authoring tool